# Automated Data Observatory

An open call to build better data spaces, observatories, or make existing data programs futureproof

Reprex is a Dutch startup that grew out of the prestigious `Yes!Delft AI+Blockchain Lab` where we have validated our `automated data observatory` concept. In Creative Europe's *JUMP European Music Market Accelerator*, we applied our idea to our beachhead market, music, where the European Commission was already planning to launch a new observatory. In 2022, we won a Horizon Europe Research and Innovation Action grant to build a prototype of a European music observatory; in 2023, our platform has been hosting several Music Moves Europe *MusicAIRE* projects.

Our product is a 2020s reimagination of the data **observatory** concept. Historically, it has been a location used for observing terrestrial, marine, or celestial events. In the last 30 years, many social, humanities, and economic observatories have been established to provide a consistent and permanent data and knowledge recording point; we have not found anyone with significant research automation capacity or who deploys trustworthy AI. Our partners can deploy a new observatory or refurbish an existing one that offers higher data quality, better data ecosystem growth, lower data collection costs and more modern services than existing observatories.
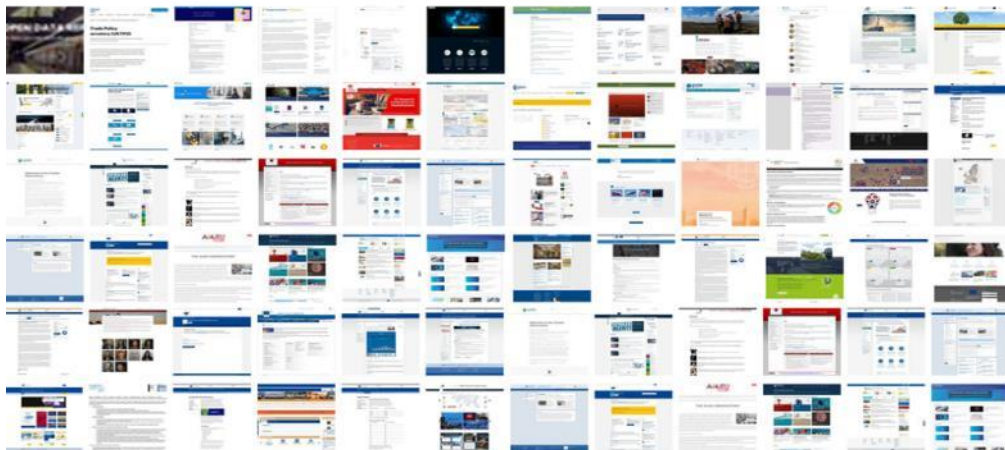
Various UN and OECD bodies, and particularly the European Union support or maintain more than 60 data observatories, or permanent data collection and dissemination points. They are often financed from the Horizon Europe programme.

About 3-5 new observatories go functional every year, and 1-2 become discontinued. Their services, data quality and quantity vary; however, in more than 95% of the cases, they receive significant public funding. They offer benefits for all parties in the knowledge triangle: academia, civil society, and innovative business. We want to bring the creation of a new observatory more affordable to our research and innovation partners and use this opportunity to introduce them to further reproducible research applications that can significantly increase the competitiveness and impact of their business processes in scientific, policy or business research.

Automated Data Observatory

Embrace the future of open data

- **Semantic web technologies** for exponentially increasing return on data investments: we organise our partner's data collection with data linking technology that enables a decentralised, secure, and inexpensive way to integrate small domain-specific databases into a big data space.

- **Taking the most out of open data**: we are creating and continuously improving an R language software ecosystem that re-processes hard-to-use open data, such as data from environmental satellites, microdata repositories of survey archives and statistical providers, anonymised government data into statistical quality datasets, as if they came from the Eurostat website–because we use the same statistical production principles as statistical authorities. *See our blogpost* Open Data - The New Gold Without the Rush.

- **Statistical innovation for exchanging privately held and governmental statistical microdata**, following the recent best practices approved by the UN and EU statistical bodies for experimental big data statistical products. *Check out our pilot program in Slovakia, which aims to be the first in-production application of the recent UN/Eurostat innovation of combining privately held data with official statistical production in the cultural and creative sectors. We aim to renew the satellite accounts of the Slovak Republic with a music and film account and show how privately-held registers and accounts can be safely coordinated with official statistical registers and survey frames (Antal 2023).*

- **Taking open science a step further** than the five-star recommendations of the Horizon Europe program: our observatories optimise data dissemination and place your deliverables directly into global library systems and global repositories. Our observatories are equipped with journal/book publishing automated workflows and can publish reports and yearbooks at ease. We also measure your impact and prove that we take it further than anything you have seen.

- **Open Policy Analysis Guidelines**: we are the first adopters of the OPA Guidelines, originally developed at Berkeley to adopt open science transparency criteria for evidence-based policy research. The advantage of OPA is greater visibility and, quicker policy impact, and better stakeholder engagement.

- **Better data governance and stewardship**: our observatory concept was developed over a decade to resolve the problems of connecting valuable private music datasets with conflicting interests and cutting through the boundaries of legal entities and countries. See you at 18th edition of IDCC24, the *International Digital Curation Conference (IDCC24), where we will present our in a global conference.*

In March, the project signed *Memorandum of Understanding on utilizing the Open Policy Analysis results of the OpenMuse Research and Innovation Consortium in the context of Slovak cultural and creative industries and sectors' public policies* (Ministerstvo kultúry SR and Open Music Europe 2023) with the Slovak Ministry of Culture to form an open policy collaboration based on the OPA Guidelines. This MoU allowed us to initiate a series of ongoing policy stakeholder consultations in the Slovak Republic with an increasing number of relevant bodies, providing very clear exploitation pathways (in the policy field) and maximising potential policy impact.

Automated Data Observatory

Join the Digital Music Observatory

The `Digital Music Observatory` is a prototype of an open, decentralised, reproducible research-supported data observatory that follows the *Feasibility Study for establishing a European Music Observatory*. High-quality indicators and databases are being developed, consulted with the industry, produced and added to the Digital Music Observatory web resource in various public and private projects. The platform is being developed by Reprex, and the four main Data Pillars by the Open Music Europe project. It's offspring, Open Collections is geared towards GLAM (galleries, libraries, archives, museums) and works with microdata, such as inventory books, catalogues, survey databases.

- For `policy researchers`, we can provide high-quality indicators that cannot be found elsewhere. Our statistical production workflow and quality assurance are similar to those of the European Statistics Code of Practice and the Quality Assurance Framework of the European Statistical System (with necessary modifications and simplification as we do open statistics and we are not part of any member state's statistical authority.) More ▸▸

- For `music businesses` and their national organisations, music information centers, libraries we provide software-as-service solutions to improve your datasets and databases. We also welcome case studies in Creative Europe or Horizon Europe projects. More ▸▸

- For `academic researchers`, we provide data and a publishing platform where your research (already from a blogpost or working paper stage) receive higher visibility and impact than it normally would. If you are not supported by a university press publisher, we provide you all the infrastructure, DOI and ISBN identifiers to make your work visible in global repositories, library catalogues, and even place your work in libraries. More ▸▸

- For `data curators and stewards` we provide best practices for research data management and data stewardship. Our entire music observatory is ready for replication, and we are already working on a `Global Green Film Observatory` and a `Global Pollinators Observatory` to place our data-to-policy pipeline, research automation tools and data infrastructure into new domains and contexts. We can provide your projects and organisations with the best available research data management, measured impacts, and complete digital publishing solutions. Our automated data observatory can provide your project *Data Management Plan* with a DCAT-API standard data catalogue (ready to be published on the European Open Data Portal, far beyond the requirements of Horizon Europe or national data management requirements), or we can take over the entire data stewardship of your project and manage your DMP from the onset of the project. (Current requirement is from the 3rd month of the project, and in some cases, from day 1.) More ▸▸

If you want to use the `Digital Music Observatory` as a digital dissemination platform, a data resource, a publishing resource, a knowledge center for your project, please join our stakeholder network.

Automated Research Data Management Task (Domain Agnostic)

Our `automated data observatory` is itself a reproducible service that can power a data space or a data observatory. It is equipped with various reproducible research, digital publishing and knowledge engineering tools that can enrich any existing data collaborations or PPPs or jump-start a new one. If you start new research projects, our technology is ideal for creating a meaningfully higher impact with your research data management.

Automated Data Observatory

Our product is being developed in itself in a Horizon Europe project based on our minimal viable product validated in the *Yes!Delft AI+Blockchain Lab Product/Market Fit* and the *JUMP* European accelerator program. As it is built from open-source components (mainly developed by Reprex and its data science partners), its various components are on TLR Levels 4-8.

- Curate and reconcile ontologies, taxonomies, and shared authority control files, to collect, enrich and integrate data from heterogeneous knowledge sources (for example, the European Environmental Agency data warehouse, the Copernicus data warehouse, primary survey sources.)

- Beyond the state-of-the-art harvesting data from ecological and socio-economic data sources, with BeeSage, we create a knowledge model to incorporate data collected from potentially hundreds of Research Hives, thus using actual beehives and their `Digital Hive Twins` as part of a cloud-based research infrastructure that is compatible with satellite surveyance of the area of the beehives, the meteorological, hydrological and socio-economic data collections of the immediate neighbourhood of them. For example, we package the data of the ResearchHives from `North Friesland` a small statistical region with business data, survey data from the same area; North Frieslands's meteorological, hydrological, and pollution data; and satellite images of leaf coverage and other essential indicators overlaying the `North Friesland` NUTS3 area.

- Use our data-to-policy pipeline and R software packages to create tidy datasets in a data catalogue for internal use in the consortium and for reuse. On tidy datasets, we mean W3C standard, future-proof datasets that are optimised to be imported into relational and graph databases (for repeating or semantic applications) and, as stand-alone datasets, can be opened in spreadsheet or statistical software packages without further data wrangling. By a data catalogue, we mean an internal consortium data catalogue following the DCAT-API application standard of the European Open Data Portal and Eurostat.

- Use our dissemination model to ensure that all public datasets, documents, figures immediately find their way directly to the Open Data Portal, and in the OpenAIRE and other relevant open science repositories, and into various library catalogues globally. Our digital data distribution boosts the scientific and community outreach of the consortium.

- For Consortium partners not supported by a University Press publisher, we provide a `Your Observatory` ISBN and DOI identifiers for all public deliverables for maximum scientific impact and findability. (Read, for example, 2 pages on the `Global Pollinator Observatory`: pdf)

- Reprex's automated data observatory components, including a peer-reviewed and functional open-source survey data harmonisation software and environmental and ecological impact assessment software, are on TLR Levels 7-9. A new observatory must be partly defined during the proposal writing process by the researchers and partially delivered in the early stage of the project. Overall, Reprex has a proven technology to provide a research infrastructure, with many elements ready to be deployed in days. However, new research functionality must be defined and added, so a new observatory can be placed to TLR Level 3 or 4 which is ideal for Horizon Europe or Creative Europe Innovation Lab funding.

If you already have a task for Research Data Management, our task can contribute to this task with a standardised data catalogue that fulfils the requirements set by the Horizon Europe program and the EU Open Data Portal in the EU Publications Office. If you do not already have such a task assigned, our task can create your project's mandatory, five-star, human-readable and machine-actionable data management plan.

Automated Data Observatory

Antal, Daniel. 2023. 'Pilot Program for Novel Music Industry Statistical Indicators in the Slovak Republic'. Zenodo. https://doi.org/10.5281/zenodo.8399254.

Ministerstvo kultúry SR, and Open Music Europe. 2023. 'Memorandum o porozumení o využití výsledkov analýz otvorených politík v kontexte slovenského kultúrneho a kreatívneho priemyslu a sektorových verejných politík v spolupráci s konzorciom pre výskum a inovácie s názvom OpenMuse. [Memorandum of Understanding on utilizing the Open Policy Analysis results of the OpenMuse Research and Innovation Consortium in the context of Slovak cultural and creative industries and sectors' public policies]'. https://www.crz.gov.sk/zmluva/7645338/.

Open Music Europe. 2023. 'Open Music Europe (OpenMusE) – An Open, Scalable, Data-to-Policy Pipeline for European Music Ecosystems'. https://doi.org/10.3030/101095295.